

2D Relative Pose and Scale Estimation with Monocular Cameras and Ranging

Chen Zhu^{*}, Gabriele Giorgi[†], Christoph Günther^{†,*}

^{*}Institute for Communications and Navigation, Technical University of Munich, Munich, Germany

[†]Institute of Communications and Navigation, German Aerospace Center (DLR), Oberpfaffenhofen-Wessling, Germany

Corresponding author:

Chen Zhu
Technical University of Munich (TUM)
Institute for Communications and Navigation (NAV)
Theresienstr. 90
80333 Munich
Germany
Telephone: +49 89 289 23435
Fax: +49 89 289 23490
Email: chen.zhu@tum.de

This work has been performed in the framework of the project VaMEx-CoSMiC (Valles Marineris Explorer-Cooperative Swarm Navigation, Mission and Control). The project VaMEx-CoSMiC is supported by the Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag, grant 50NA1521 administered by DLR Space Administration.

Keywords— visual SLAM, visual navigation, relative positioning, cooperative positioning, sensor fusion

ABSTRACT

Cooperative swarms of robots equipped with cameras are robust against failures, and can explore GNSS (Global Navigation Satellite System)-denied environments efficiently. Applying VSLAM (Visual Simultaneous Localization and Mapping) techniques, vehicles can estimate their trajectories and simultaneously reconstruct the map of the environment using visual cues. Due to constraints on payload size, weight, and costs, many VSLAM applications must be based on a single camera. The associated monocular estimation of the trajectory and map is ambiguous by a scale factor. This work shows that by exploiting sparse range measurements between a pair of dynamic rovers in planar motion, the correct scale factors of both cameras and the relative position, as well as the relative attitude between the rovers, can be estimated. Neither images nor feature vectors are required to be transmitted over the communication channel for the proposed method, which is a significant advantage in practice.

INTRODUCTION

Autonomous robotic platforms are utilized in the exploration of extreme environments, e.g., extraterrestrial exploration or catastrophe rescues. In order to increase the system robustness against hazards in the missions, e.g., strike during landing, and to improve the exploration efficiency, we propose to use a robotic swarm including multiple autonomous units such as multicopters and ground rovers [1] [2]. Autonomous navigation of the swarm elements often relies on several sensors such as mobile receivers, Inertial Measurement Units (IMUs), laser scanners and, most substantially, cameras [3]. Due to constraints on size, weight, accommodation and costs in swarm elements, monocular cameras are used instead of stereo rigs in most cases. VSLAM techniques using monocular cameras have been developed in recent years to estimate the trajectory of vehicles and to simultaneously reconstruct the map of the environment. Klein and Murray developed the Parallel Tracking and Mapping (PTAM) algorithm [4], which divides the tracking and mapping into separate threads to accelerate the computation. Engel, Schöps and Cremers proposed a large scale dense SLAM algorithm using monocular cameras [5], which minimizes the photometric error instead of the feature reprojection error for reducing the computational costs and improving the performance. Another state-of-the-art approach is ORB-SLAM from Mur-Artal, Montiel and Tardós [6]. The method utilizes ORB (Oriented FAST and Rotated BRIEF) features [7] and a novel keyframe-based graph structure, to provide a robust real-time monocular SLAM solution even in large scale scenarios and relatively low frame rate. However, all these algorithms estimate the motion only up to a global scale.

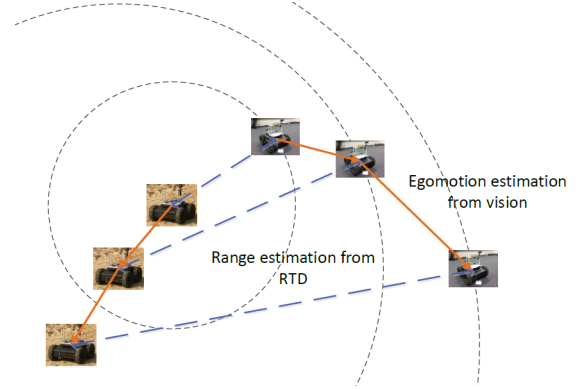


Figure 1: Two dynamic rovers with ranging measurements

A number of approaches have been considered for resolving the global scale ambiguity. Most of them use IMUs, see for example Nützi et al. [8] and Abeywardena et al. [9]. However, the inherent drift of IMUs is prone to introducing estimation biases. As a consequence, onboard laser range finder (LIDAR) is used by a number of authors, e.g., Zhang et al. [10], Bodensteiner et al. [11] and Caselitz et al. [12], to solve the scale problem of monocular VSLAM. This leads to excellent results. However, in many application scenarios based on monocular cameras, the utilization of LIDARs is limited by the constraints on weight, size, and costs. Therefore, we developed a method for estimating the global scales of a pair of dynamic rovers in planar motion, using sparse range measurements on a single ranging link. In the case of a swarm of robots, these measurements could be performed between any pair of swarm elements [13]. Strictly, the algorithm developed in this paper does not depend on the method of ranging. It can be adapted without restrictions from radio-frequency-based ranging to other sources of ranging measurements, e.g., radar or lidar. Fig. 1 shows a scenario of two dynamic rovers equipped with monocular cameras and a ranging link between them. By exploiting the cooperation between the pair of vehicles, the scale problem in VSLAM of both monocular cameras can be solved with the proposed method. Additionally, the scale estimation problem couples with the estimation of the rovers' initial relative position and attitude. As a consequence, the relative pose between the two rovers can be obtained within the same framework.

This manuscript is structured as follows. First, the system model and a brief introduction of motion estimation in monocular-camera-based VSLAM are introduced. Then, a method for the scale and relative pose estimation of two cooperative rovers using monocular cameras and sparse range measurements is proposed. Finally, several simulation results are produced to test the method's performance under different rover trajectories and measurement noise, and conclusions are drawn upon an analysis of the estimation outcomes.

SYSTEM MODEL AND MOTION ESTIMATION USING MONOCULAR CAMERAS

The measurement scenario addressed in this work is shown in Fig. 1. Two cooperative rovers equipped with a monocular camera and a ranging device, e.g., a wireless radio receiver, execute SLAM tasks on the ground. The motion of the vehicle is constrained to be planar. We define a navigation frame (N) as a fixed coordinate frame for each rover with its origin at the starting location of rover. The navigation frame of each rover is related to the world reference frame by a specific transformation dependent on the initial position and attitude of vehicle. Moreover, we use (k) to express the camera's local coordinate frame at keyframe k , which varies as the camera moves. Let $\vec{c}_{[k]}^{(W)} \in \mathbb{R}^2$ be the position of the robot in world frame (W) at time k . In the remainder of this paper, we use a superscript with parentheses (\cdot) to denote the coordinate frame in which the vector is represented. Vectors such as $\vec{c} \in \mathbb{R}^2$ with geometric meanings are written with an arrow. Time, denoted with square brackets $[\cdot]$, is measured in keyframes, i.e., the time reference instances in which both the range measurements and the trajectory estimation are available. The homogeneous coordinates in the extended Euclidean plane are written as $\tilde{r} \in \mathbb{P}^2$. In addition, the origin of the body frame is defined at the position of the ranging sensor. Since the relative pose between the camera and the ranging sensor can be obtained by calibration, the body frame and camera frame are not distinguished. This assumption does not affect the validity of the algorithm if the body is assumed to be rigid.

The range measurements can be obtained by using pilot signals for synchronization. If the clock on the transmitter and receiver sides are precisely synchronized, the range can be estimated using time of arrival (ToA) measurements. If a satisfactory synchronization cannot be achieved, round-trip-delay (RTD) techniques can be implemented to eliminate the impact of the clock offset. The precision of the range measurements is constrained by their Cramér-Rao lower bound [14]. The details of ranging using RTD for navigation purposes are discussed in [13].

In the proposed scheme, the rovers have basic communication capabilities so that one of them can transmit its local estimated trajectory $\{\vec{c}_{1,[k]}^{(N_1)}\}$ to the other one. The trajectory is estimated by a VSLAM algorithm in the navigation frame of the rover, i.e., the fixed reference frame taking the starting location as the origin and the initial heading direction as the y-axis. Our method does not require transmission of extracted feature vectors or the local maps, so the data throughput requirement is significantly low. A radio-based system with both ranging and communication capabilities for robotic swarms is proposed by Zhang et. al. in [15].

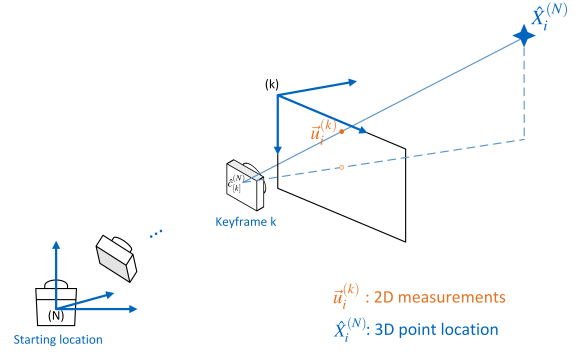


Figure 2: Camera projection model in navigation frame

To obtain the trajectory in navigation frame $\{\vec{c}_{[k]}^{(N)}\}$, the following steps of monocular-camera-based motion estimation are essential. Generally, the transformation between two coordinate frames (P) and (Q) follows

$$\vec{X}^{(Q)} = R_{(P \rightarrow Q)} \vec{X}^{(P)} + \vec{t}_{(P \rightarrow Q)}, \quad (1)$$

where $\vec{X}^{(P)}$ and $\vec{X}^{(Q)}$ denote the coordinates of an arbitrary 3D point $\vec{X} \in \mathbb{R}^3$ expressed in the corresponding (P) and (Q) frames, $R_{(P \rightarrow Q)} \in \mathbf{SO}(3)$ denotes the orthonormal rotation matrix, and $\vec{t}_{(P \rightarrow Q)}$ denotes the translation vector from the origin of (P) to the origin of (Q).

According to perspective projection, a visible point with 3D coordinates in the navigation frame $\vec{X}_i^{(N)} \in \mathbb{R}^3$ is projected to a two-dimensional (2D) point $\vec{u}_i^{(k)}$ in the measurement set $\Omega_{[k]}$ at k -th keyframe as

$$\vec{u}_i^{(k)} = \pi(\vec{X}_i^{(N)}, \vec{c}_{[k]}^{(N)}, R_{(N \rightarrow k)}) \in \Omega_{[k]} \subset \mathbb{R}^2. \quad (2)$$

$\Omega_{[k]}$ is the set consisting of the 2D coordinates of all the points of interest on the image plane. In feature-based approaches, e.g., in [4] and [6], the measurement space Ω is continuous, whereas in direct methods such as [5] and [16] it is a discrete set, i.e. the set of all the pixels. For the widely applied pinhole camera model with lens distortion correction [17], the projection can be simply denoted in homogenous coordinates as

$$\tilde{u}_i^{(k)} = K P_{[k]} \tilde{X}_i^{(N)}, \quad (3)$$

where K denotes the camera intrinsic matrix, and $P_{[k]}$ the extrinsic projection matrix at time k . Figure 2 illustrates the pinhole camera model for the projection at k -th keyframe. In planar motion case,

$$P_{[k]} = R_{(N \rightarrow k)} \begin{bmatrix} I_3 & -\vec{c}_{[k]}^{(N)} \\ 0 & 1 \end{bmatrix}, \quad (4)$$

where I_3 denotes the three-dimensional identity matrix.

By tracking features in consecutive image sequences, the essential matrix $E_{(k \rightarrow k+1)}$ can be estimated using the epipolar geometry constraint:

$$(K^{-1} \tilde{u}_i^{(k+1)})^T E_{(k \rightarrow k+1)} (K^{-1} \tilde{u}_i^{(k)}) = 0. \quad (5)$$

The essential matrix can be decomposed into a rotation $R_{(k \rightarrow k+1)}$ and a unit vector of translation $\vec{e}_{(k \rightarrow k+1)} \in \mathbb{R}^3$ as: $E_{(k \rightarrow k+1)} = [\vec{e}_{(k \rightarrow k+1)}]_{\times} R_{(k \rightarrow k+1)}$, where $[\cdot]_{\times}$ denotes the 3×3 skew symmetric matrix built as

$$[\begin{matrix} e_1 \\ e_2 \\ e_3 \end{matrix}]_{\times} = \begin{bmatrix} 0 & -e_3 & e_2 \\ e_3 & 0 & -e_1 \\ -e_2 & e_1 & 0 \end{bmatrix}. \quad (6)$$

However, the distance of the translation is not obtainable according to the epipolar constraint, due to the invariance of Eq. (5) to the scaling of $E_{(k \rightarrow k+1)}$.

The translation in true scale is related to the monocular estimation by

$$\vec{t}_{(k \rightarrow k+1)} = s_g l_{(k \rightarrow k+1)} \vec{e}_{(k \rightarrow k+1)}. \quad (7)$$

In this equation $l_{(k \rightarrow k+1)} \vec{e}_{(k \rightarrow k+1)}$ is the estimated translation from monocular vision, in which $l_{(k \rightarrow k+1)} \in \mathbb{R}^+$ denotes the estimated norm of the translation from time k to $k+1$, and $\vec{e}_{(k \rightarrow k+1)}$ denotes the direction of the motion. $s_g \in \mathbb{R}^+$ is the true global scale in the world frame, which cannot be obtained in the monocular-only case [18]. The relative scale between two translations can be extracted. Without loss of generality, one can assume $l_{(1 \rightarrow 2)} = 1$. The 3D coordinates of the tracked points can be estimated by triangulation to build a local map. A local optimization, e.g., bundle adjustment [19], shall be applied using the estimated motion to initialize the tracking thread of the S-LAM algorithm. Then, the positions at the following time instances can be obtained by minimizing the re-projection residual (photometric residual in direct method cases)

$$\hat{c}_{[k]}^{(N)} = \arg \min_{\substack{\vec{c}_{[k]}^{(N)} \\ \vec{u}_i^{(k)} \in \Omega_{[k]}}} \sum \left\| \pi(\vec{X}_i^{(N)}, \vec{c}_{[k]}^{(N)}) - \vec{u}_i^{(k)} \right\|_{\Sigma^{-1}}, \quad (8)$$

where Σ is the measurements covariance matrix.

SCALE AND RELATIVE POSE ESTIMATION EXPLOITING SPARSE RANGE MEASUREMENTS

Without any other anchor point with known absolute position, one can only estimate the position and attitude of the cameras with respect to a selected point in the navigation frame. We choose the initial position of the camera projection center of rover 2 as the coordinate reference system's origin, and the camera's principal axis as the y-axis. Fig. 3 illustrates the reference system and the geometry of the two rovers. The initial position and attitude of the two rovers can be expressed in the reference frame as

$$\vec{c}_{1,[1]}^{(W)} = r_1 R(\alpha) [1, 0]^T, \quad R_{(N_1 \rightarrow W)} = R(\alpha + \theta - \frac{\pi}{2}). \quad (9)$$

$$\vec{c}_{2,[1]}^{(W)} = [0, 0]^T, \quad R_{(N_2 \rightarrow W)} = I_2, \quad (10)$$

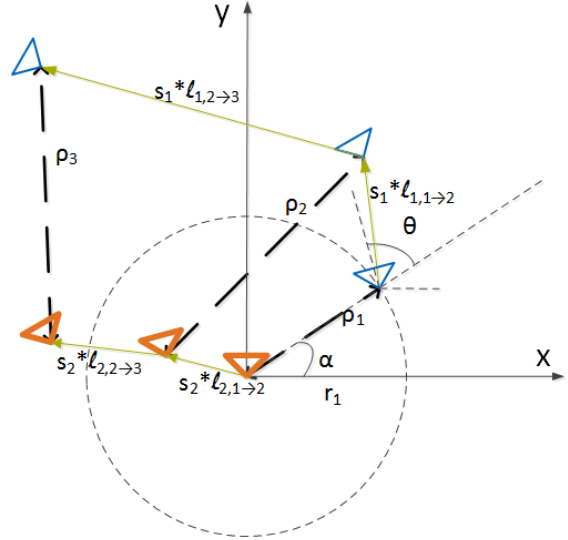


Figure 3: Reference system and the geometry of the two rovers

where I_2 denotes the two-dimensional identity matrix, and $R(\cdot) \in \mathbf{SO}(2)$ denotes a 2D rotation matrix.

Using the images from the monocular cameras, the egomotion of the two rovers in their navigation frames can be independently estimated up-to-scale as $\{\vec{c}_{1,[k]}^{(N_1)}\}$ and $\{\vec{c}_{2,[k]}^{(N_2)}\}$. In the common reference frame (W), the position of the two rovers at k -th keyframe can be expressed as

$$\vec{c}_{1,[k]}^{(W)} = s_{g1} R_{(N_1 \rightarrow W)} \vec{c}_{1,[k]}^{(N_1)} + \vec{c}_{1,[1]}^{(W)} \quad (11)$$

$$\vec{c}_{2,[k]}^{(W)} = s_{g2} \vec{c}_{2,[k]}^{(N_2)} \quad (12)$$

Although the monocular camera itself can only estimate the motion with a scale ambiguity, with the additional help of a sparse set of noisy range measurements $\{\rho_k\}$, where

$$\rho_k = \left\| \vec{c}_{1,[k]}^{(W)} - \vec{c}_{2,[k]}^{(W)} \right\| + \eta_k, \quad (13)$$

a method for estimating the scale factors s_{g1}, s_{g2} can be devised by exploiting consecutive ranging measurements at keyframes. The true range between the two rovers at time k is

$$G_k(s_{g1}, s_{g2}, \alpha, \theta, r_1) = \left\| \vec{c}_{1,[k]}^{(W)} - \vec{c}_{2,[k]}^{(W)} \right\| = \left\| s_{g1} R(\alpha + \theta - \frac{\pi}{2}) \vec{c}_{1,[k]}^{(N_1)} + r_1 R(\alpha) [1, 0]^T - s_{g2} \vec{c}_{2,[k]}^{(N_2)} \right\|, \quad (14)$$

which is determined by the rover trajectories in navigation frames and 5 unknown scalar parameters: the scale factors $s_{g1}, s_{g2} \in \mathbb{R}^+$, the polar angle $\alpha \in [0, 2\pi)$, the attitude angle $\theta \in [0, 2\pi)$, and the initial distance $r_1 \in \mathbb{R}^+$. These five unknown parameters are stacked in a vector $\xi = [s_{g1}, s_{g2}, \alpha, \theta, r_1]^T$.

By utilizing the communication functionality of the radio link between the two rovers, rover 1 can transmit its estimated motion (up-to-scale) to rover 2. Rover 2 serves as the master that obtains both trajectory estimates in local frames. Neither images nor feature vectors are required to be transmitted over the communication channel for this method, which is a significant advantage in practice. By using the available set of range measurements along with the local trajectory estimates, the unknown parameters can be estimated by minimizing

$$\hat{\xi} = \arg \min_{\xi} \|\rho - G(\xi)\|_{Q^{-1}}^2, \quad \text{s.t. } B\xi > 0, \quad (15)$$

with vectors $\rho = [\rho_1, \rho_2, \dots, \rho_K]^T$ and $G(\xi) = [G_1(\xi), G_2(\xi), \dots, G_K(\xi)]^T$.

$B = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$ is a selection matrix used to impose the positiveness of both scales and the initial distance. Q is the covariance matrix that characterizes the ranging measurement noise $\eta = [\eta_1, \eta_2, \dots, \eta_K]^T$. The Cramér-Rao lower bound of the range estimation can be used as an approximation when the covariance calculation is unavailable. If the ranging noise is uncorrelated across time, Q is a diagonal matrix.

Due to the bounded search space and the presence of several local minima, it is challenging to solve the nonlinear inequality constrained optimization in Eq. (15). However, not all minima violating the constraints represent erroneous solution, due to the symmetric properties of the objective function. According to Eq. (14), the norm G_k is invariant if the vector $\tilde{c}_{1,[k]}^{(W)} - \tilde{c}_{2,[k]}^{(W)}$ is reversed in direction. Consequently, for any parameter vector ξ , the value of the object function is invariant to the following parameter change:

$$\begin{aligned} & G_k(s_{g1}, s_{g2}, \alpha, \theta, r_1) \\ &= G_k(-s_{g1}, s_{g2}, \alpha, \theta + \pi, r_1) \\ &= G_k(-s_{g1}, -s_{g2}, \alpha + \pi, \theta + \pi, r_1) \\ &= G_k(-s_{g1}, -s_{g2}, \alpha, \theta, -r_1) \\ &= G_k(-s_{g1}, s_{g2}, \alpha + \pi, \theta, -r_1) \\ &= G_k(s_{g1}, -s_{g2}, \alpha + \pi, \theta, r_1) \\ &= G_k(s_{g1}, -s_{g2}, \alpha, \theta + \pi, -r_1) \\ &= G_k(s_{g1}, s_{g2}, \alpha + \pi, \theta + \pi, -r_1). \end{aligned} \quad (16)$$

As a result, due to the symmetry property of the cost function, any solution of the corresponding unconstrained problem can be transformed to a valid solution which satisfies $B\xi > 0$. Therefore, we can obtain the estimates of the parameters by solving the unconstrained problem and transform the results using Table 1, if any of s_{g1}, s_{g2} or r_1 has negative value from the unconstrained optimizer.

The nonlinear optimization problem (15) can be linearized to an unconstrained linearized least-squares problem

$$\hat{\xi} = \arg \min_{\xi} \|\rho - J(\xi)\xi\|_{Q^{-1}}^2, \quad (17)$$

with Jacobian matrix

$$J(\xi) = \begin{bmatrix} \frac{\partial G_1(\xi)}{\partial s_{g1}} & \frac{\partial G_1(\xi)}{\partial s_{g2}} & \frac{\partial G_1(\xi)}{\partial \alpha} & \frac{\partial G_1(\xi)}{\partial \theta} & \frac{\partial G_1(\xi)}{\partial r_1} \\ \frac{\partial G_2(\xi)}{\partial s_{g1}} & \frac{\partial G_2(\xi)}{\partial s_{g2}} & \frac{\partial G_2(\xi)}{\partial \alpha} & \frac{\partial G_2(\xi)}{\partial \theta} & \frac{\partial G_2(\xi)}{\partial r_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial G_K(\xi)}{\partial s_{g1}} & \frac{\partial G_K(\xi)}{\partial s_{g2}} & \frac{\partial G_K(\xi)}{\partial \alpha} & \frac{\partial G_K(\xi)}{\partial \theta} & \frac{\partial G_K(\xi)}{\partial r_1} \end{bmatrix}.$$

The optimization (17) can be solved iteratively as

$$\hat{\xi}_{i+1} = \hat{\xi}_i + \left(J^T(\hat{\xi}_i) Q^{-1} J(\hat{\xi}_i) \right)^{-1} J^T(\hat{\xi}_i) Q^{-1} (\rho - G(\hat{\xi}_i)). \quad (18)$$

The scales of both trajectories, as well as the initial relative position and attitude between the two rovers, are thus estimated. By combining these estimates with the trajectories previously obtained in the navigation frames, the relative pose at any given keyframe k can be extracted. As a distributed system, the master rover can transmit the estimation results to the other one using the available communication channel.

In order to solve the problem in Eq. (17), $K \geq 5$ range measurements are required. Due to the high nonlinearity of the objective function, the Levenberg-Marquardt algorithm [20] is applied, instead of a Gauss-Newton approach [21], in order to exploit its better global minimization capabilities. In addition, the initialization of the optimization is crucial due to the presence of a number of local minima. Although a suboptimal solution may have similar residual as the global minimum, the estimated parameters can be far away from the true value, leading to a wrong scale or pose. A precise approximation (ρ_1) of the initial range r_1 is generally available thanks to the high accuracy of ranging measurements. Initializing the scale factors may be more difficult, but the convergence to the correct value proves to be significantly insensitive to the initial conditions of the global minimization problem, provided that the selected keyframes are sufficiently spaced (which should be significantly larger than the ranging noise). The estimation of the polar angle α and the attitude angle θ presents larger difficulties. Fortunately, the parameters to be estimated are constants and in most cases they do not need to be updated at high frequency. Hence a serial search for the proper initialization of the two angles is feasible. It is remarkable that if the relative position between the two rovers can be estimated by other methods, e.g., using ranging measurements from the swarm network in [13], the polar angle α could be precisely initialized. As a result, the search space would even reduce to a one-dimensional set.

Table 1: Transformation on the results from unconstrained optimization.

If			Transformation				
$\hat{s}_{g1} > 0$	$\hat{s}_{g2} < 0$	$\hat{r}_1 > 0$	$\hat{s}_{g1} \leftarrow -\hat{s}_{g1}$	$\hat{s}_{g2} \leftarrow -\hat{s}_{g2}$	$\hat{\alpha} \leftarrow \hat{\alpha} + \pi$	$\hat{\theta} \leftarrow \hat{\theta}$	$\hat{r}_1 \leftarrow -\hat{r}_1$
$\hat{s}_{g1} > 0$	$\hat{s}_{g2} < 0$	$\hat{r}_1 < 0$	$\hat{s}_{g1} \leftarrow -\hat{s}_{g1}$	$\hat{s}_{g2} \leftarrow -\hat{s}_{g2}$	$\hat{\alpha} \leftarrow \hat{\alpha}$	$\hat{\theta} \leftarrow \hat{\theta} + \pi$	$\hat{r}_1 \leftarrow -\hat{r}_1$
$\hat{s}_{g1} > 0$	$\hat{s}_{g2} > 0$	$\hat{r}_1 < 0$	$\hat{s}_{g1} \leftarrow -\hat{s}_{g1}$	$\hat{s}_{g2} \leftarrow -\hat{s}_{g2}$	$\hat{\alpha} \leftarrow \hat{\alpha} + \pi$	$\hat{\theta} \leftarrow \hat{\theta} + \pi$	$\hat{r}_1 \leftarrow -\hat{r}_1$
$\hat{s}_{g1} < 0$	$\hat{s}_{g2} > 0$	$\hat{r}_1 > 0$	$\hat{s}_{g1} \leftarrow -\hat{s}_{g1}$	$\hat{s}_{g2} \leftarrow -\hat{s}_{g2}$	$\hat{\alpha} \leftarrow \hat{\alpha}$	$\hat{\theta} \leftarrow \hat{\theta} + \pi$	$\hat{r}_1 \leftarrow -\hat{r}_1$
$\hat{s}_{g1} < 0$	$\hat{s}_{g2} < 0$	$\hat{r}_1 > 0$	$\hat{s}_{g1} \leftarrow -\hat{s}_{g1}$	$\hat{s}_{g2} \leftarrow -\hat{s}_{g2}$	$\hat{\alpha} \leftarrow \hat{\alpha} + \pi$	$\hat{\theta} \leftarrow \hat{\theta} + \pi$	$\hat{r}_1 \leftarrow -\hat{r}_1$
$\hat{s}_{g1} < 0$	$\hat{s}_{g2} < 0$	$\hat{r}_1 < 0$	$\hat{s}_{g1} \leftarrow -\hat{s}_{g1}$	$\hat{s}_{g2} \leftarrow -\hat{s}_{g2}$	$\hat{\alpha} \leftarrow \hat{\alpha}$	$\hat{\theta} \leftarrow \hat{\theta}$	$\hat{r}_1 \leftarrow -\hat{r}_1$
$\hat{s}_{g1} < 0$	$\hat{s}_{g2} > 0$	$\hat{r}_1 < 0$	$\hat{s}_{g1} \leftarrow -\hat{s}_{g1}$	$\hat{s}_{g2} \leftarrow -\hat{s}_{g2}$	$\hat{\alpha} \leftarrow \hat{\alpha} + \pi$	$\hat{\theta} \leftarrow \hat{\theta}$	$\hat{r}_1 \leftarrow -\hat{r}_1$

SIMULATION RESULTS

We test the proposed method on multiple trajectories using simulation data with Gaussian additive noise. The trajectories are generated with random walk processes as accelerations, starting from static locations with random relative position and attitude. In the simulation, two noise sources are considered: ranging noise – with standard deviation σ_p – and estimation errors on the relative translation vectors – with standard deviation σ_t . In order to simulate a realistic scenario, the error on the trajectory estimation is added on all the translation estimates instead of on positions, i.e., the error accumulates over time.

Fig. 4 shows the relation between the scale estimation accuracy of both cameras and the ranging noise σ_p . The error of scale is calculated as the root-mean-square error (RMSE) of the estimated scale factor s_{g1} and s_{g2} with 200 simulation runs, each under 5 randomly generated trajectories with 500 keyframes and random initial relative poses. The true values of the scale factors are between 0.3 to 0.5 in the simulation, generated according to the random walk of acceleration. To better decouple the impact of error sources, the translation error is set to 0 in this simulation. From the curves it can be concluded that the scale estimation is still reliable even with 40 to 50 [cm] ranging accuracy, given the motion estimation is sufficiently precise. On the other hand, the change of scale estimation error with respect to the increase of the translation error σ_t is shown in Fig. 5 for a ranging error-free scenario. It can be seen that if the ranging estimation is error-free, the global scale factor can be recovered with high accuracy for propagated translation errors as large as 20 [cm] between consecutive keyframes.

In real scenarios, the ranging noise and translation error always exists at the same time. For the trajectories shown in Fig. 6, the root-mean-square error of the parameter estimation under different noise levels is shown in Table 2. All the RMSE are calculated with ten repetitive runs with independent noise. The trajectory of rover 2, i.e., the master node, is plotted in orange with thicker lines and rover 1 is in blue. Fig. 7 shows the first 30 frames to illustrate the initial relative geometry more clearly. In the serial search of initial values of the polar angle α and attitude angle θ ,

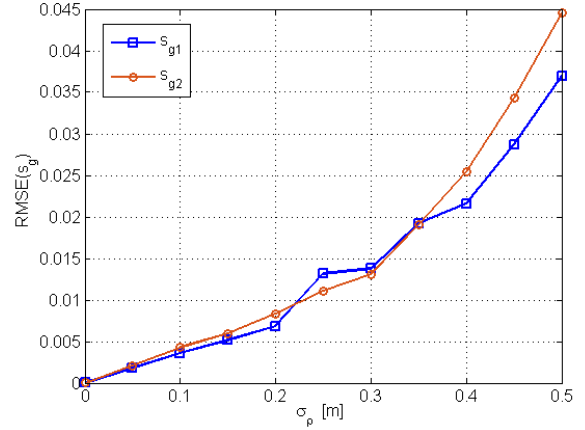


Figure 4: RMSE of scale estimation with respect to ranging noise.

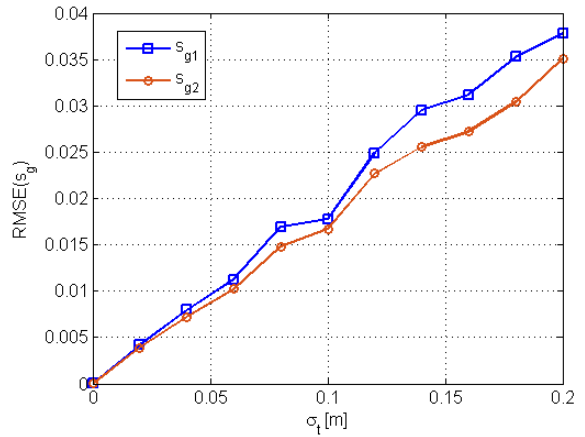


Figure 5: RMSE of scale estimation with respect to translation error.

Table 2: Estimation error of scales and pose parameters in Scenario #1.

σ_t	1 [cm]	1 [cm]	3 [cm]	5 [cm]	5 [cm]
σ_p	1 [cm]	10 [cm]	10 [cm]	10 [cm]	20 [cm]
$RMSE(s_{g1})$	0.0016	0.0049	0.0120	0.0129	0.0086
$RMSE(s_{g2})$	0.0015	0.0045	0.0127	0.0116	0.0067
$RMSE(\alpha)$ [deg]	3.3893	3.3426	8.2128	6.9004	8.8601
$RMSE(\theta)$ [deg]	0.6539	1.8069	4.2246	8.5225	6.2905
$RMSE(r_1)$ [m]	0.0171	0.0301	0.0596	0.1620	0.0716

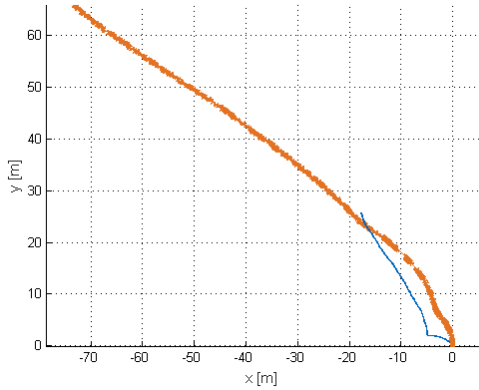


Figure 6: Trajectories of the two rovers in Scenario #1. The master rover trajectory is plotted in orange with thicker lines.

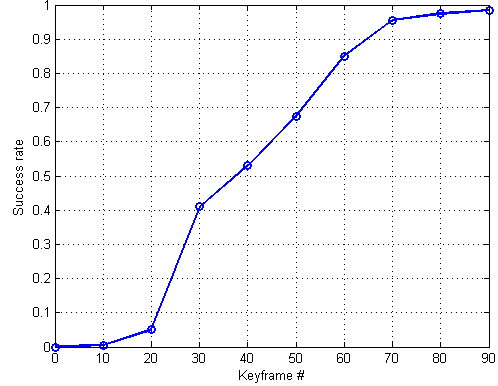


Figure 8: Success rate with respect to number of keyframes.

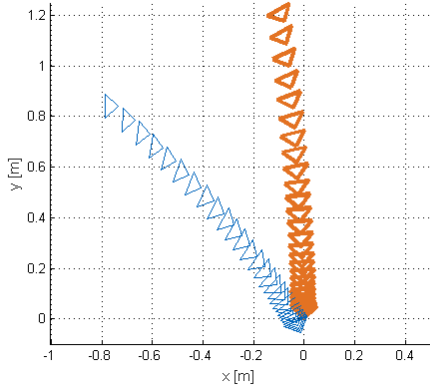


Figure 7: First 30 frames of the two rovers in Scenario #1.

the grid size is set to 10 degrees in the simulation. It can be concluded from the results that in this scenario the estimation scheme converges well even for 5 [cm] translation error and 20 [cm] ranging noise. The scale factors in both trajectories can be accurately estimated, with errors not exceeding 3%. An improvement in the angular estimation precision can be obtained by setting a higher density of serial search values in the initialization of the non-linear optimization. Using our C++ based implementation on a desktop computer, the execution time of the proposed method is shown in Table 3.

Fig. 8 shows the performance of the algorithm using only a few keyframes. The success rate of the algorithm is evaluated in the following way: the algorithm is applied to 200 runs with independently generated noise samples for

the ranging and motion estimation. The error in $\hat{\xi}$ is evaluated after 10, 20, ..., 90 keyframes. Specifically, whenever the error in $\hat{s}_{g1}, \hat{s}_{g2}$ and \hat{r}_1 is less than 10%, and the error of the estimated angles $\hat{\alpha}, \hat{\theta}$ is less than 10 degrees, the outcome of the algorithm is considered as a success. This success rate is represented in Fig. 8 with respect to the number of keyframes used. In the simulation, the initialization steps for α and θ are both set to be 60 [deg]. The standard deviation of the translation noise is set as 3 [cm], while the ranging noise is 10 [cm]. It can be concluded from the curve that the method can achieve a success rate higher than 95% after 70 keyframes (with travel distance around 20 [m]).

Other scenarios are also simulated to test the performance of the proposed method in different motion geometries. The trajectories of the rovers in various scenarios are shown in Fig. 9 and the corresponding estimation results are given in Table 4. Scenario #4 is a special situation that the motion of both rovers are constrained to be linear and are with opposite headings. It can be concluded that the method performs well in various scenarios with different geometries. A key factor that affects the precision of the estimation is the magnitude of the simulated motion. If the change of distance between the two rovers is comparable to the ranging noise, the measurement noise would be dominant in the estimation.

Table 3: Elapsed running time for C++ based implementation.

Elapsed time [s]:	2.218	0.530	0.255	0.123
Total number of initialization:	1296	288	144	64
α initialization step [deg]:	10	10	30	45
θ initialization step [deg]:	10	45	30	45

CONCLUSION AND FUTURE WORK

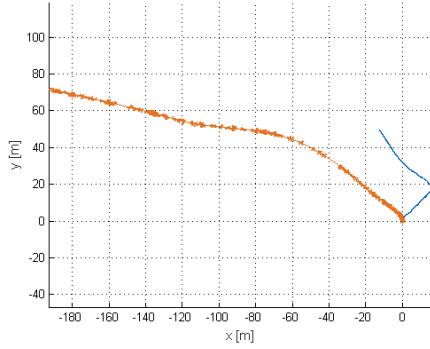
In many vision applications, a single camera is preferred over a stereo rig due to weight and cost constraints. However, the global scale is not recoverable in monocular vision. We propose an algorithm to resolve the global scale ambiguity in monocular VSLAM for a pair of cameras mounted on two rovers moving independently on a plane. By exploiting range measurements between the two rovers, the correct scales of the egomotions are estimated. At the same time, the relative position and attitude can be obtained. The algorithm was successfully tested on a number of simulated datasets with various geometries and noise patterns. Based on the proposed method, the global scale and relative pose estimation can be extended to multiple rovers, provided that multiple access wireless radio channels are used. As a result, the formation of a robotic swarm can be estimated. The work can also be extended to three-dimensional motion scenarios. However, three parameters are required to parameterize any vehicle attitude in 3D cases. The convergence to global optima is much more challenging.

REFERENCES

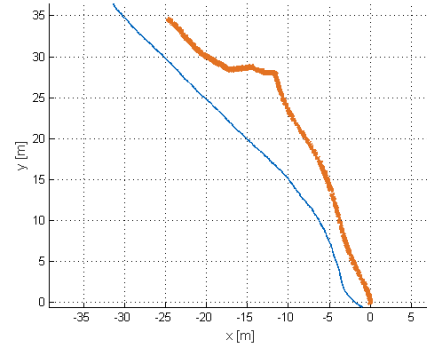
- [1] Y. Mohan and S. G. Ponnambalam, "An extensive review of research in swarm robotics," in *Nature Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*, Dec. 2009, pp. 140–145.
- [2] S. Sand, S. Zhang, M. Mühlegg, G. Falconi, C. Zhu, T. Krüger, and S. Nowak, "Swarm exploration and navigation on Mars," in *International Conference on Localization and GNSS, Torino, Italy*, 2013.
- [3] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers," *Journal of Field Robotics*, vol. 24, no. 3, pp. 169–186, 2007.
- [4] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, Nov 2007, pp. 225–234.
- [5] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Computer Vision - ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, vol. 8690, pp. 834–849.
- [6] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.
- [7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: an efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2564–2571.
- [8] G. Nützi, S. Weiss, D. Scaramuzza, and R. Siegwart, "Fusion of imu and vision for absolute scale estimation in monocular slam," *Journal of Intelligent and Robotic Systems*, vol. 61, no. 1-4, pp. 287–299, 2011.
- [9] D. Abeywardena, Z. Wang, S. Kodagoda, and G. Disanayake, "Visual-inertial fusion for quadrotor micro air vehicles with improved scale observability," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, May 2013, pp. 3148–3153.
- [10] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low-drift, robust, and fast," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2174–2181.
- [11] C. Bodensteiner, W. Hübner, K. Jüngling, P. Solbrig, and M. Arens, "Monocular camera trajectory optimization using lidar data," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2018–2025.
- [12] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d lidar maps," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 1926–1931.
- [13] E. Staudinger, S. Zhang, A. Dammann, and C. Zhu, "Towards a radio-based swarm navigation system on mars - key technologies and performance assessment," in *Wireless for Space and Extreme Environments (WiSEE), 2014 IEEE International Conference on*, Oct 2014, pp. 1–7.
- [14] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

Table 4: Estimation error of parameters in various scenarios.

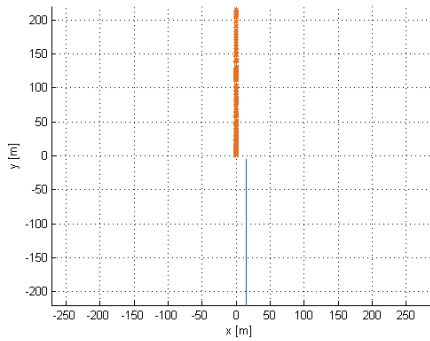
Scenario #	2	3	4	5
σ_t	5 [cm]	5 [cm]	3 [cm]	1 [cm]
σ_p	20 [cm]	10 [cm]	10 [cm]	5 [cm]
$RMSE(s_{g1})$	0.0052	0.0030	0.0127	0.0009
$RMSE(s_{g2})$	0.0018	0.0017	0.0046	0.0033
$RMSE(\alpha)$ [deg]	2.9416	9.7634	0.0005	5.8405
$RMSE(\theta)$ [deg]	3.5057	6.0490	0.0004	0.7336
$RMSE(r_1)$ [m]	0.1337	0.0717	0.0761	0.0205



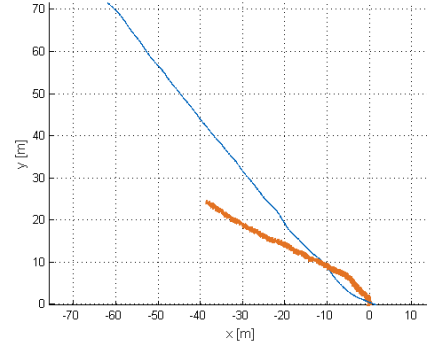
(a) Scenario #2



(b) Scenario #3



(c) Scenario #4



(d) Scenario #5

Figure 9: Trajectories of the two rovers in various scenarios.

- [15] S. Zhang, S. Sand, R. Raulefs, and E. Staudinger, "Self-organized hybrid channel access method for an interleaved RTD-based swarm navigation system," in *Workshop on Positioning, Navigation and Communication, Dresden, Germany*, 2013.
- [16] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, March 2018.
- [17] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [18] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *Robotics & Automation Magazine, IEEE*, vol. 18, no. 4, pp. 80–92, 2011.
- [19] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment: a modern synthesis," in *Vision algorithms: theory and practice*. Springer, 1999, pp. 298–372.
- [20] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.
- [21] Y. Wang, "Gauss-newton method," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 4, pp. 415–420, 2012.